SD-Pose: Structural Discrepancy Aware Category-Level 6D Object Pose Estimation

Guowei Li[1,2], Dongchen Zhu[1,2], Guanghui Zhang[1], Wenjun Shi[1], Tianyu Zhang[1,2],Xiaolin Zhang[1,2,3,4,5], and Jiamao Li [1,2,3*]

[1]Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Xiongan Institute of Innovation, Xiongan, 071700, China
[4]University of Science and Technology of China, Hefei, Anhui, 230027, China
[5]ShanghaiTech University, Shanghai 201210, china
jmli@mail.sim.ac.cn

Abstract

Category-level 6D object pose estimation aims to predict the full pose and size information for previously unseen instances from known categories, which is an essential portion of robot grasping and augmented reality. However, the core challenge of this task still is the enormous shape variation within each category. With regard to the challenge, we propose a novel framework SD-Pose, which utilizes the instance-category structural discrepancy and the potential geometric-semantic association to enhance the exploration of the intra-class shape information. Specifically, an information exchange augmentation (IEA) module is introduced to supplement the instance-category structural information by their structural discrepancy, thus facilitating the enhanced geometric information to contain both the character of instance shape and the commonality of category structure. For complementing the deficiencies of structural information adaptively, a semantic dynamic fusion (SDF) module is further designed to fuse semantic and geometric features. Finally, the proposed SD-Pose framework equipped with the IEA and SDF modules hierarchically supplements instance-category structural information in a stacked manner and achieves state-of-the-art performance on the CAMERA25 and REAL275 datasets.

1. Introduction

Accurately estimating the 6D pose of an object is a quite crucial task in computer vision, which is widely employed in real-world applications such as 3D scene understanding [35], robotic grasping [9], virtual reality [1], and augmented reality [25, 34]. 6D object pose estimation includes instance-level and category-level methods. So far, instance-level 6D pose estimation works [19, 29, 22, 27, 38, 17, 16] have made considerable progress. However, as an accurate CAD model is usually required during the training and inference, instance-level methods can only deal with a few objects or just a single instance, severely limiting their practical application in the real world. For breaking instance-level constraints, category-level 6D pose estimation proposes to predict the complete pose information for previously unseen objects from known classes [39]. In this paper, we focus on the category-level 6D pose estimation task, which

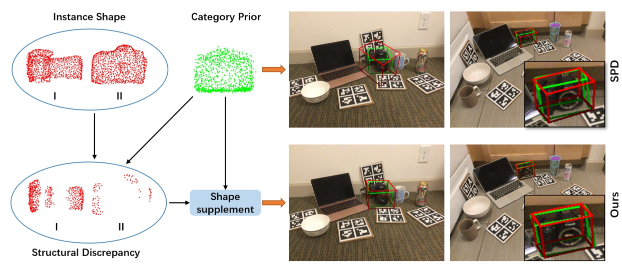


Figure 1: Comparing pose estimation results between the SPD [36] and our structural discrepancy supplement method. The category prior and two camera instances have different structural discrepancies. The red and green lines are prediction results and ground truth, respectively.

is a more general assignment due to does not rely on the instance CAD model.

Currently, the critical challenge of the category-level task is still the extreme shape variation within each class [33, 31, 32]. To overcome the problem of intra-class variation, Wang *et al*. [39] introduce Normalized Object Coordinate Space(NOCS)—a share canonical representation for all possible object instances within a category. Some works then [39, 2, 36, 20] learn the RGB-D features of each object instance to reconstruct the CAD model of the object instance with the same size and orientation in NOCS. However, such a reconstruction process lacks the implicit representation of shape variations, limiting pose estimation performance.

Concerning this problem, SPD [36] proposes generating a category prior for each class and deforming it to reconstruct the NOCS model of the object instance. Although the SPD has achieved sound effects, such a fixed category prior can only reflect the fuzzy structure information and cannot capture local structure changes for each instance. Especially when the structural discrepancy between the category prior and the instance is enormous, it becomes difficult to reconstruct an accurate object model, severely affecting the pose estimation performance. Fortunately, the category prior can be supplemented by structural discrepancy derived from the instance-category geometry relationship to better match the instance model. As shown in Figure 1, each camera instance and category prior have a distinct difference in structure. Compared to SPD [36], our method performs more acceptable by utilizing the structural discrepancy to supplement the category prior. Particularly when the structural discrepancy is enormous, the improvement is more prominent. Furthermore, since the structural discrepancies of the category prior and corresponding diverse instances are distinct, our method is able to accommodate previously unseen instances of various shapes, dramatically increasing the generalization of our method.

In this paper, we propose a novel category-level pose estimation framework SD-Pose, which leverages the structural discrepancy between instance and category prior to enhancing the learning of intra-class shape information. Furthermore, considering the inaccuracy NOCS model of reconstructed instance caused by category prior ambiguity, we recommend combining additional semantic information following [13, 41, 11]. Specifically, we further design a Semantic Dynamic Fusion (SDF) module to dynamically adjust the semantic information through the geometry relationship and fuse it with enhanced category prior to adaptively supplementing the lack of structural information. In summary, our main contributions are as follows:

- An Information Exchange Augmentation (IEA) module is introduced to guide the category prior more reasonable suit the instance geometry by utilizing instance-category structural discrepancy to enhance the respective geometric features.
- For complementing structural information deficiencies adaptively, a Semantic Dynamic Fusion (SDF) module is further designed to fuse category prior and instance semantic features with a dynamic adjustment according to the instance-category structural relationship.
- Based on stacking multiple IEA and SDF modules, a novel category-level pose estimation framework SD-Pose is proposed to learn intra-class shape variations by exploiting the instance-category structural relationship. Our SD-Pose achieves a state-of-the-art performance on CAMERA25 and REAL275 datasets.

2. Related Works

2.1. Instance-Level 6D Object Pose Estimation

In instance-level tasks, the object CAD model is known at the training and inference stages, which can be roughly classified into three different approaches: template-based, correspondence-based, and voting-based. Template-based methods [18, 26, 30] need to find the template most similar to the object image or point cloud from the template sets labeled with the ground truth 6D pose, which is a part-to-all coarse registration problem. The correspondence-based method aims to find the correspondence between the observed object and its complete CAD model. For the correspondence between 2D and 3D [27, 29, 30], the pose is obtained by solving a PnP problem [21]. As for the correspondence between 3D and 3D [6, 7], the pose is calculated by the least-squares method. The voting-based method can be divided into direct voting and indirect voting. Directly voting [38, 16] returns a 6D pose and confidence score at each position and then selects the most reliable pose information as the final result. Indirect voting [27, 17] first selects key point positions through RANSAC [10] voting and then calculates the 6D pose of the object according to the correspondence between key points.

2.2. Category-Level 6D Object Pose Estimation

Category-level tasks aim to predict pose information for the previously unseen object instance, which is formally introduced in [39]. Wang *et al*. [39] use a normalized object coordinate space (NOCS) to represent all objects in the same class. Then they reconstruct the instance CAD model in NOCS and adopt the Umeyama [37] algorithm to calculate the pose with the NOCS model and observed points. Due to the huge intra-class shape variation, some later methods pay more attention to the geometric information of the object. CASS [2] obtains a canonical shape space by learning. DualPoseNet [24] utilizes a dual-stream

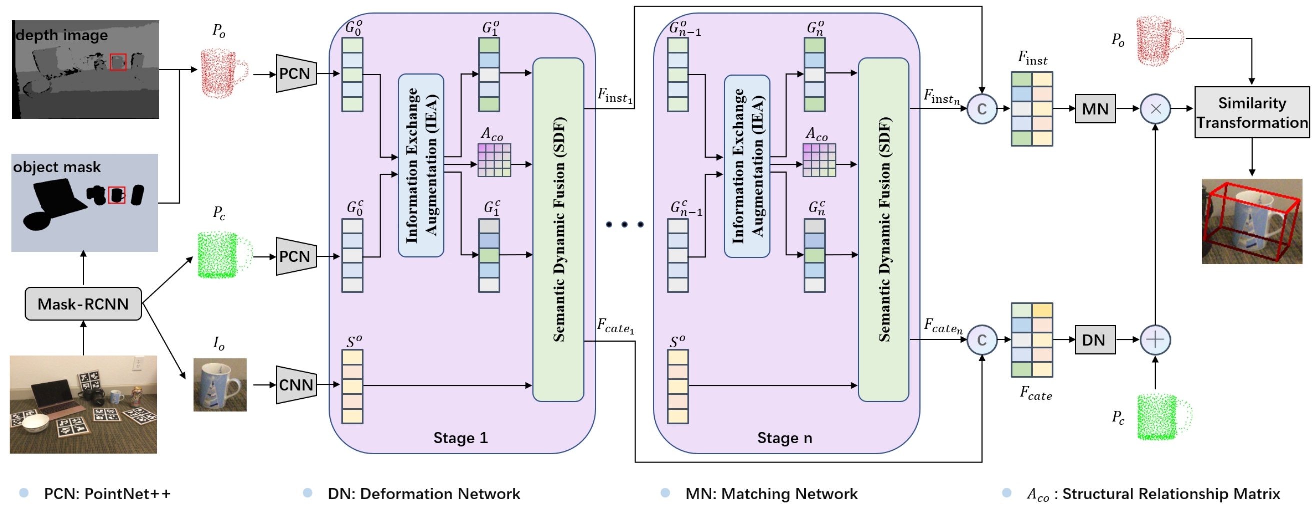


Figure 2: An overview of our proposed SD-Pose framework. Firstly, taking image patch I_o, observed point cloud P_o, and category prior P_c as inputs, instance semantic features S^o, instance geometry features G^o, and category geometry features G^c are obtained by features extracted module. Then Information Exchange Augmentation (IEA) module is utilized to supplement geometry features G^o and G^c. After that, a Semantic Dynamic Fusion (SDF) module is employed to fuse semantic and geometry features. By stacking multiple IEA and SDF modules, the final instance features F_{inst} and category features F_{cate} are generated. Next, we reconstruct the instance NOCS model and establish the correspondence between the observed point and the NOCS model. Finally, the 6D pose is recovered by estimating a similarity transformation. Here G^o_{n-1} and G^c_{n-1} are the output of IEA module of stage $n-1$.

network to explicitly and implicitly encode pose information and uses pose consistency to optimize the pose. FS-Net [8] decodes orientation information through a decoupled rotation mechanism. Do-Net [23] exploits symmetry for pose optimization. Although these methods improve performance, they can not explicitly harness the structural relationship between pose and point cloud. Other methods instead utilize a category prior to reconstruct a 3D model of the NOCS space. In exploring category priors, CR-Net [40] explores the complex and informative relations among instance RGB image, instance point cloud, and category prior to advance representation learning. In addition, SGPA [5] leverages instance-category structural similarity to dynamically adapt the prior to the observed object, which is most relevant work as ours. Different from it, in this work, we explore the structural discrepancy between instance and category prior to learn intra-class shape change, which reflects the unique geometric appearance of each instance more directly and effectively. Compared to utilizing structural similar, a more accurate instance NOCS model can be rebuilt after the category prior is supplemented by the structural discrepancy.

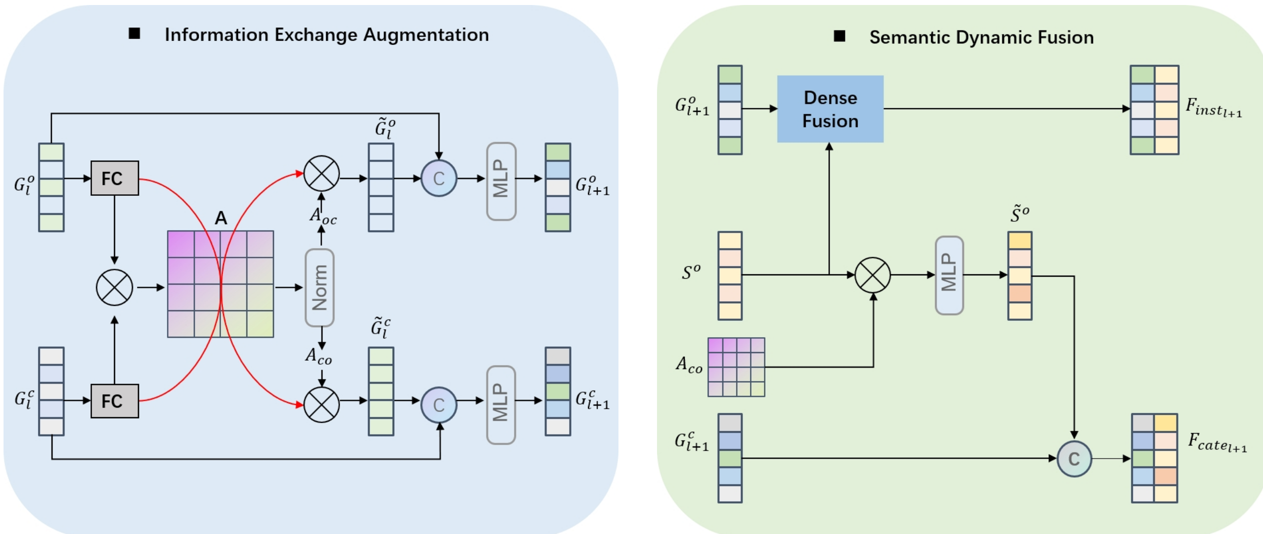
3. Methodology

Problem Formulation. Given an RGB-D image, our task is to estimate 6D pose of and 3D size of the object from its partially observed point cloud. We represent the 6D object pose as a rigid transformation matrix $[R|t] \in \mathcal{SE}(3)$ consisting of a rotation $R \in \mathcal{SO}(3)$ and a translation $t \in \mathbb{R}^3$ matrix. The 3D size of the object is described as $s \in \mathbb{R}^3$

Pre-processing Stage. Following SPD [36], we first employ an off-the-shelf object detection network(*e.g.* Mask-RCNN [14]) to obtain RGB image patches of observed objects $I_o \in \mathbb{R}^{h \times w \times 3}$, where (h, w) is the image block size. The observed point cloud $P_o \in \mathbb{R}^{n_o \times 3}$ comes from depth channel, where n_o is the number of instance point clouds. $P_c \in \mathbb{R}^{n_c \times 3}$ is the category prior corresponding to the observed object, where n_c is the number of category point clouds.

3.1. Overview

Here we give an overview of our SD-Pose, as in Figure 2. Taking I_o, P_o and P_c as inputs, we first use a feature extraction module to obtain instance semantic features S^o, instance geometric features G^o and category geometric features G^c (Section 3.2). Leveraging the structural relationship matrix $A \in \mathbb{R}^{n_o \times n_r}$ of G^c_0 and G^c_0, the IEA module supplements the original geometric features by implicitly encoding structural discrepancy features to acquire enhanced instance geometric features G^o_1 and category geometric features G^c_1 (Section 3.3). Afterwards, S^o, G^o_1, G^c_1 and A will be fed into the SDF module to proceed seman-



(a) IEA (b) SDF

Figure 3: The structure of IEA and SDF module in l-th stage. (a) IEA takes instance geometry features G_l^o and category geometry features G_l^c as inputs to learn geometry relation matrix A, thus calculating the structural discrepancy features according to A and enhancing their original geometry features. (b) SDF takes instance semantic features S^o, enhanced instance geometry features G_{l+1}^o , category geometry features G_{l+1}^c and geometry relationship matrix A_{co} as inputs. For S^o and G_{l+1}^o, We adapt a pixel-wise dense fusion method [38] to obtain instance features $F_{inst_{l+1}}$. Then, we fuse G_{l+1}^c and $\tilde{S}^o$ dynamically adjusted by structural relation A_{co} to obtain category feature $F_{cate_{l+1}}$. $\bigotimes$ denotes matrix multiply.

tic and geometric features fusion to obtain instance features F_{inst_1} and category features F_{cat_1} (Section 3.4). To ensure sufficient interaction of instance and category structural information, the IEA and SDF are embedded into the framework in a stacked manner. The features of each stage are stitched to get the final instance features F_{inst} and category features F_{cate}. Later, following SPD [36], Later, following SPD [36], a deformation network is utilized to reconstruct the instance NOCS model by deforming the category prior P_c. Moreover, a matching network is adopted to match the reconstructed model with the observed point cloud P_o. Finally, the correspondence-based algorithm [37] is applied to estimate pose parameters (Section 3.5).

3.2. Feature Extraction

The feature extraction module is first employed to learn semantic and geometric features. Specifically, the image patch I_o is processed by the PSPNet [42] with the backbone of ResNet-18 [15] to obtain semantic features $S^o \in \mathbb{R}^{n_o \times d_c}$, which is a point-wise pixel features corresponding to the observed point cloud encouraged by Densefusion [38]. For point cloud P_o and P_c, we employ the PointNet++ [28] to extract instance geometric features $G_0^o \in \mathbb{R}^{n_o \times d_c}$ and category geometric features $G_0^c \in \mathbb{R}^{n_c \times d_c}$ respectively.

3.3. Information Exchange Augmentation

Our IEA module aims to learn the structural relationship between instance point clouds and category prior, which can assist in constructing their structural discrepancy information at the feature level. It utilizes features of structural discrepancy to supplement the original geometric features, making the enhanced geometric features include the unique individuality of instance structure and general commonality of category prior. On the one hand, due to complementing peculiarity of the instance structural, the enhanced category geometry features can reconstruct a more accurate instance NOCS model. On the other hand, instance geometric features add category shape commonality, thereby promoting the rebuilt correspondence matrix better associate the observed point cloud with the NOCS model. Moreover, since the geometric discrepancy between category prior and different instances under the same class are distinct, our method is able to accommodate previously unseen instances of various shapes, dramatically increasing the generalization of our method.

The structure of the IEA module is shown in Figure 3a. As Givens the instance geometric features G_l^o and category geometric features G_l^c of the l-th stage, we project them to the feature subspace of the same dimension by a Fully Connected layer and then adopt the matrix multiplication operation to obtain the structural relationship matrix

$A \in \mathbb{R}^{n_o \times n_r}$:

$$A = FC(G_l^o) \times FC(G_l^c)^T \tag{1}$$

Following the normalization method [12] designed specifically for point cloud attention map, A further is normalized in two different dimensions respectively to acquire weight matrices A_{oc} and A_{co}:

$$\begin{aligned} a_{ij}^o &= \frac{e^{A_{ij}}}{\sum_{k=1}^{n_o} e^{A_{kj}}}, A_{ij}^{oc} = \frac{a_{ij}^o}{\sum_{k=1}^{n_r} a_{ik}} \\ a_{ij}^c &= \frac{e^{A_{ij}}}{\sum_{k=1}^{n_r} e^{A_{ik}}}, A_{ij}^{co} = \frac{a_{ij}^c}{\sum_{k=1}^{n_o} a_{kj}^c} \end{aligned} \tag{2}$$

After that, the geometric projection features perform weighted summation by the corresponding structural weight matrices to gain structural discrepancy features $\tilde{G}_l^c$ and $\tilde{G}_l^o$:

$$\tilde{G}_l^c = (A_{co})^T \times FC(G_l^o), \tilde{G}_l^o = A_{oc} \times FC(G_l^c) \tag{3}$$

Finally, We joint the original geometric features and structural discrepancies features by exploiting the Multi-layer Perceptron (MLP) function to obtain enhanced geometric features G_{l+1}^o and G_{l+1}^c:

$$\begin{aligned} G_{l+1}^o &= MLP(Concat(G_l^o, \tilde{G}_l^o) \\ G_{l+1}^c &= MLP(Concat(G_l^c, \tilde{G}_l^c)) \end{aligned} \tag{4}$$

3.4. Semantic Dynamic Fusion

As shown in Figure 4, the input observed point cloud, which is obtained using Mask-RCNN [14] segmentation results rather than ground truth, probably contains some outliers. When the influence of these outliers is transmitted to the category prior, it will theoretically have a negative impact on the reconstruction accuracy of the NOCS model, and lead to a deviation in the correspondence between the observed point cloud and instance NOCS model. Fortunately, the additional semantic information can help alleviate these problems. Inspired by [11, 13, 41], we design the SDF module, which seeks to reduce the influence of noise points by fusion sufficiently the geometric and semantic information, improving the robustness of the network to noise points.

Figure 3b illustrates the SDF module. For the fusion of geometric features G_{l+1}^o and semantic features S^o of the instance from different modalities, the key lies in how to integrate cross-modal features [3, 4, 38] effectively. Inspired by Densefusion [38], a point-wise fusion module is achieved to explore the intrinsic mapping between data sources by adopting a pixel-level correspondence strategy. The fused features are output as $F_{inst_{l+1}}$.

As for the fusion of category geometric features G_{l+1}^c and instance semantic features S^o, the pixel-level fusion

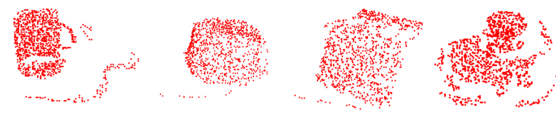


Figure 4: Observed point clouds of different instances are obtained by Mask-RCNN segmentation. Each instance contains some noise points.

method cannot be used directly because they belong to different individuals, that is, there is no pixel-level correspondence. Intuitively, following the general idea of feature fusion, we only concatenate and fuse them through an MLP function to obtain $F_{cate_{l+1}}$. We call it semantic immediate fusion (SIF):

$$F_{cate_{l+1}} = MLP(concat(S^o, G_{l+1}^c)) \tag{5}$$

Although the designed SIF can improve performance via absorbing semantic information immediately, it is still ill-considered for the cross-individual problem. Hence, we further design a semantic dynamic fusion (SDF) module, which dynamically adjusts instance semantic features S^o according to the instance-category structural relationship matrix A_{co} and combines with G_{l+1}^c to obtain the category features $F_{cate_{l+1}}$. It can be formulated as

$$\begin{aligned} \tilde{S}^o &= A_{co} \times S^o \\ F_{cate_{l+1}} &= MLP(Concat(G_{l+1}^c, \tilde{S}^o)) \end{aligned} \tag{6}$$

We prefer the latter method because G_{l+1}^c and S^o belong to different individuals with a specific domain diversity. Dynamically adjusting semantic information through the structural relationship matrix A^{co} may be significantly aware of individual differences and improve the generality of the network to unseen object instances. The experimental results (Table 3) further demonstrate that the latter fusion strategy can achieve better performance.

3.5. Pose Estimation

We separately joint output of each stage to obtain final instance features F_{inst} and category features F_{cate}:

$$\begin{aligned} F_{inst} &= Concat(F_{inst_1}, \cdots, F_{inst_n}) \\ F_{cate} &= Concat(F_{cate_1}, \cdots, F_{cate_n}) \end{aligned} \tag{7}$$

After obtaining F_{inst} and F_{cate}, we estimate the pose following SPD [36]. Specifically, a deformation network first is utilized to regress a deformation field point by point $D \in \mathbb{R}^{N_r \times 3}$ and deform P_c to reconstruct the instance NOCS standard model:

$$\hat{P}_N = P_c + D = P_c + \mathcal{F}_d(F_{inst}, F_{cate}) \tag{8}$$

Table 1: Comparisons with other methods on CAMERA25 and REAL275 datasets.

Method	CAMERA25						REAL275					
	IoU50	IoU75	5^o2cm	5^o5cm	10^o2cm	10^o5cm	IoU50	IoU75	5^o2cm	5^o5cm	10^o2cm	10^o5cm
NOCS [39]	83.9	69.5	32.3	40.9	48.2	64.6	78.1	30.1	7.2	10.0	13.8	25.2
CASS [2]	-	-	-	-	-	-	77.7	-	-	23.5	-	58.0
SPD [36]	93.2	83.1	54.3	59.0	73.3	81.5	77.3	53.2	19.3	21.4	43.2	54.1
Dual [24]	92.4	86.4	64.7	70.7	77.2	84.7	79.8	62.2	29.3	35.9	50.5	66.8
SGPA [5]	93.2	88.1	**70.7**	74.5	**82.7**	**88.4**	80.1	61.9	35.9	39.6	61.3	70.7
Ours	**93.4**	**88.3**	**70.7**	**75.6**	80.5	87.7	**83.2**	**68.2**	**37.1**	**42.0**	**62.0**	**71.2**

where $\mathcal{F}_d(\cdot)$ refers to the deformation network, $\hat{P}_N \in \mathbb{R}^{N_r \times 3}$ corresponds to the NOCS standard model of the reconstructed instance object.

We then regress a corresponding matrix $M \in \mathbb{R}^{N_o \times N_r}$ through a matching network, which relates $\hat{P}_N$ to P_o:

$$\hat{P}_o = M \times \hat{P}_N = \mathcal{F}_m(F_{inst}, F_{cate}) \times \hat{P}_N \tag{9}$$

where $\mathcal{F}_m(\cdot)$ refers to the matching network, $\hat{P}_o$ is the transformed instance point cloud model and has a point-to-point correspondence with P_o. Given $\hat{P}_o$ and P_o, the corresponding method is finally employed to estimate the 6D pose of the target.

Overall, our SD-Pose has two outputs to calculate 6D pose estimation: the point-wise deformation field D, and the correspondence matrix M. In order to train SD-Pose, we adopt the same strategy with SPD [36]. The reconstruction loss by calculating the Chamfer Distance(CD) between $\hat{P}_N$ and the ground truth NOCS model P_N to penalize D:

$$L_{cd} = \sum_{i \in P_N} \min_{j \in \hat{P}_N} ||i - j||_2^2 + \sum_{j \in \hat{P}_N} \min_{i \in P_N} ||i - j||_2^2 \tag{10}$$

Then, we constrain the distance between the predicted NOCS coordinate value x and the ground-truth one x_{gt} to supervise M.Specific detail refer to SPD [36].

$$L_{cor} = \frac{1}{N_o} \begin{cases} 5\,(x - x_{gt})^2 & |x - x_{gt}| \leq 0.1 \\ |x - x_{gt}| - 0.05 & \text{otherwise} \end{cases} \tag{11}$$

4. Experiments

4.1. Experiments Setup

Datasets. We conduct experiments on category-level benchmarks of CAMERA25 and REAL275 datasets [39]. CAMERA25 is a synthetic dataset generated by a context-aware mixed reality approach. REAL275 is a more challenging real dataset.

Evaluation Metrics. Following the widely adopted evaluation scheme [39, 36, 2], we compute the average precision of 3D Intersection Over Union (IoU) at the threshold of 50% and 75% for 3D object detection. To directly evaluate errors in rotation and translation, the average precision of $m^o ncm$ is adopted.

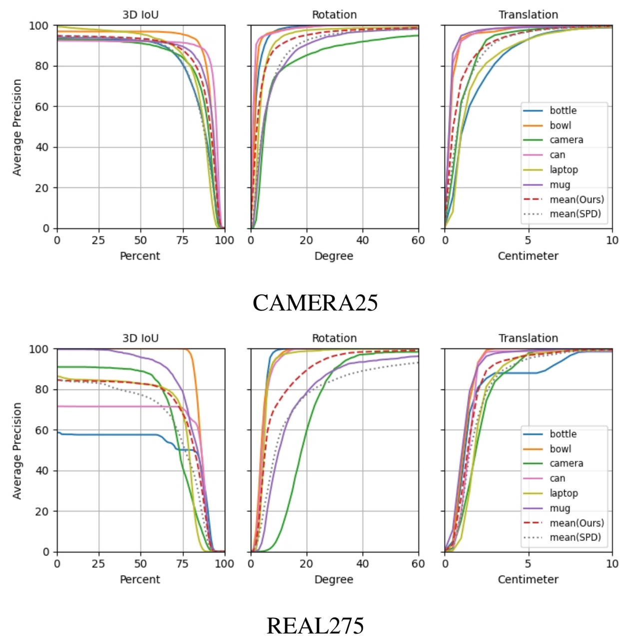


Figure 5: Average precision vs. error thresholds on CAMERA25 and REAL275 datasets.

Implementation Details. Similar to [36, 5], we decouple the instance segmentation and the subsequent pose estimation. Follow [36], we generate the instance segmentation results offline with an off-the-shelf object detector (*e.g.* Mask-RCNN [14]). After that, The target object is cropped from the RGB-D image based on the segmentation results and recover the instance point clouds utilizing camera intrinsic parameters. The image crop is resized 192×192. The number of points in the observed point cloud and category prior is downsampled to 1024. For the feature extraction module of SD-Pose, we use PSPNet [42] with backbone of ResNet-18 [15] to extract semantic features. The geometric features are acquired by a PointNet++ [28]. As for the number of stacking IEA and SDF modules, the n is set to be 2. We use an RTX 2080 Ti GPU to train SD-Pose for 50 epochs with a batch size of 32. We initially set the learning rate as 0.0001

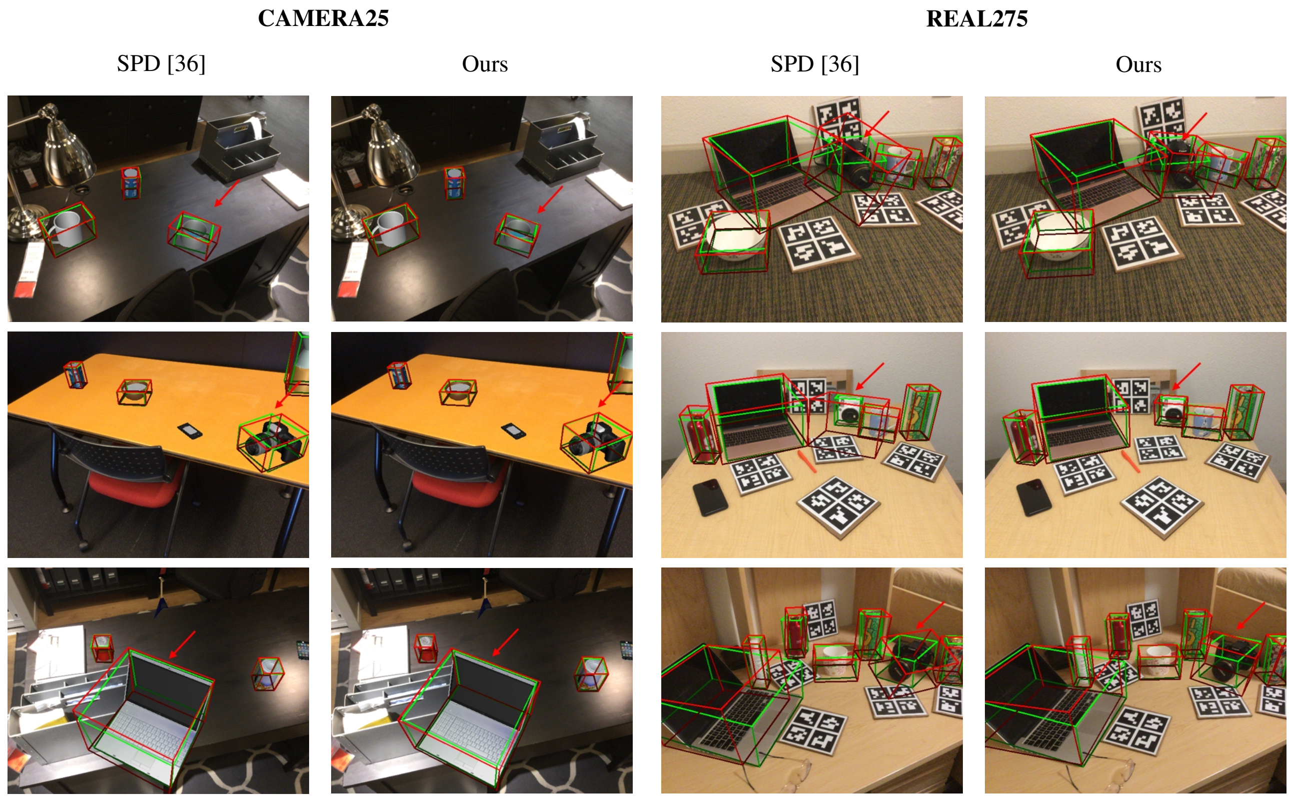


Figure 6: Qualitative comparisons between ours and SPD [36] on CAMERA25 and REAL275 datasets. We visualize the estimated 6D pose and size as the tight-oriented bounding box around the target instances. The red and green lines are prediction results and ground truth, respectively.

Table 2: Quantitative comparison of the model reconstruction accuracy in CD metric ($\times 10^{-3}$).

Method	CAMERA25						
	bottle	bowl	camera	can	laptop	mug	mean
SPD [36]	1.72	1.55	4.28	0.96	1.99	1.36	1.78
Ours	**1.29**	**1.01**	**2.70**	**0.91**	**1.12**	**1.15**	**1.36**
	REAL275						
	bottle	bowl	camera	can	laptop	mug	mean
SPD [36]	3.44	1.21	8.89	1.56	2.91	**1.02**	3.17
Ours	**1.84**	**1.02**	**5.37**	**1.38**	**1.11**	1.08	**1.96**

and halved it every 5 epochs.

4.2. Comparison with State-Of-The-Art Methods

In Table 1, we compare the proposed method with NOCS [39], CASS [2], SPD [36], Dual [24], SGPA [5]. For the synthetic CAMERA25 dataset, our method outperforms the baseline SPD on all metrics by a large margin and achieves optimal results over other methods under metrics IoU50, IoU75, and $5^o 5cm$. Besides, the indicators $10^o 2cm$ and $10^o 5cm$ are comparable to the state-of-the-art method SGPA. For the more challenging REAL275 dataset, the superiority of our method is more obvious. The proposed method significantly outperforms the current best method SGPA on all metrics, with improvements of 3.1%, 7.7%, 1.2%, 2.4%, 0.7%, 0.5%, respectively. Notably, our method realizes a more significant improvement on the real REAL275 dataset, which contains more previously unseen instances, than on the synthetic CAMERA25 dataset. It shows that our method has a good generalization. We believe this may be mainly since our method fully considers the diversity of the structural discrepancies, thus being able to accommodate previously unseen instances of various shapes.

To thoroughly verify pose estimation performance, we conduct an experimental evaluation from the perspective of model reconstruction. The Chamfer Distance of the reconstructed NOCS model with the ground truth NOCS model is computed. Comparing our method with the baseline SPD, as shown in Table 2, we can observe that the average reconstruction error of our method is lower than SPD in both datasets. It proves again that our method can achieve better pose estimation performance.

Furthermore, Figure 6 shows a qualitative comparison of two datasets. We can observe that our method produces a more accurate pose than SPD, especially on geometrically

Table 3: Ablation studies of key components tested on CAMERA25 and REAL275. IEA means Information Exchange Augmentation (Section 3.3). SIF means Semantic Immediate Fusion; SDF means Semantic Dynamic Fusion (Section 3.4).

ROW	IEA	SIF	SDF	CAMERA25						REAL275					
				IoU50	IoU75	5^o2cm	5^o5cm	10^o2cm	10^o5cm	IoU50	IoU75	5^o2cm	5^o5cm	10^o2cm	10^o5cm
1	-	-	-	93.1	85.1	55.1	59.7	74.4	82.1	79.9	59.5	19.2	21.2	45.9	56.6
2	✓	-	-	93.3	88.3	63.3	67.5	78.8	85.4	83.1	66.7	25.6	30.5	49.9	63.5
3	✓	✓	-	93.2	87.2	63.8	68.2	79.1	85.6	82.8	66.5	32.9	38.7	51.6	64.0
4	✓	-	✓	**93.5**	**88.4**	**64.9**	**69.1**	**80.5**	**86.6**	**83.2**	**67.0**	**34.2**	**39.4**	**53.0**	**64.6**

complex objects(*e.g*. camera category). This indicates our SD-Pose can sufficiently learn the shape change by utilizing instance-category structural discrepancy to supplement geometry information. In addition, we present a more detailed error evaluation result on two datasets in Figure 5, which further illustrates that our SD-Pose outperforms SPD in terms of 3D IoU, rotation, and translation.

4.3. Ablation Studies

In order to verify the efficacy of the critical components of our method, we conduct ablation studies for IEA and SDF modules on the CAMERA25 and REAL275, as shown in Table 3. For convenience, the n is set to be 1. The baseline is SPD [36] corresponding to **row 1**.

Effectiveness of IEA. We first verify the significance of using the IEA module, which can be figured out by comparing **row 1** and **row 2** in Table 3. Relative to results in **row 1**, the performance of all metrics in **row 2** has an overall boost. On the one hand, the category geometry features complement the unique individuality of instance structure to reconstruct a more accurate instance NOCS model. On the other hand, instance geometric features add general commonality of category prior, thereby facilitating the reconstructed correspondence matrix better associate the observed point cloud with the reconstructed model.

SIF or SDF? We also explore the importance of fusing semantic information and the impact of different fusion methods on performance. Comparing the results in **row 2**, **row 3** and **row 4**, after adding semantic information, there is a large improvement in the angle and translation evaluation, but SIF (**row 3**) slightly decrease in 3D IoU. This may be because semantic cues and category prior come from different individuals. Simple fusion (SIF) without the awareness of individual discrepancy may bring feature confliction to a certain degree. While dynamically adjusting semantic information through instance-category structural relationships can weaken this problem and achieve better results.

Since our network stacks multiple IEA and SDF modules, we also verify the impact of choosing the different n, where n takes values from 1, 2, and 3. Comparing the results in Table 4, we can conclude that $n = 2$ is the best choice. In this case, the mutual complementation of instance and category in structural information is optimal, thus generating better performance on pose estimation.

Table 4: Evaluation of SD-Pose on CAMERA25 and REAL275 benchmarks when n is set to be different values.

n	CAMERA25					
	IoU50	IoU75	5^o2cm	5^o5cm	10^o2cm	10^o5cm
1	**93.5**	**88.4**	64.9	69.1	**80.5**	86.6
2	93.4	88.3	**70.7**	**75.6**	**80.5**	**87.7**
3	93.2	87.6	70.2	75.0	80.3	**87.7**
	REAL275					
	IoU50	IoU75	5^o2cm	5^o5cm	10^o2cm	10^o5cm
1	**83.2**	67.0	34.2	39.4	53.0	64.6
2	**83.2**	**68.2**	**37.1**	**42.0**	**62.0**	**71.2**
3	82.4	66.2	36.5	41.1	57.9	68.2

5. Conclusion

In this paper, we propose a novel category-level 6D object pose estimation framework SD-Pose, which utilizes instance-category structural discrepancy and geometric-semantic potential association to enhance the learning of intra-class shape variation. Specifically, the IEA module is designed to supplement the instance-category geometry information by their structural discrepancy, thus facilitating the enhanced geometry information to contain both the character of instance shape and the commonality of category prior. Furthermore, the SDF module is further proposed to alleviate the influence of noise points by fusing category prior and instance semantic features with a dynamic adjustment. Our method achieves state-of-the-art performance on CAMERA25 and REAL275 datasets. Although we alleviate the problem of noise points by an implicit SDF module, it may be further optimized in our future work through an explicit manner (*e.g*. designing an appropriate point cloud filter).

Acknowledgement. This research was supported by National Science and Technology Major Project from Minister of Science and Technology, China(2021ZD0201403), National Natural Science Foundation of China(61873255), Shanghai Municipal Science and Technology Major Project (ZHANGJIANG LAB) under Grant 2018SHZDZX01, Youth Innovation Promotion Association, Chinese Academy of Sciences(2021233) and Shanghai Academic Research Leader(22XD1424500).

References

[1] Grigore C Burdea and Philippe Coiffet. *Virtual reality technology*. John Wiley & Sons, 2003.

[2] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020.

[3] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3051–3060, 2018.

[4] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. *Pattern Recognition*, 86:376–385, 2019.

[5] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021.

[6] Wei Chen, Jinming Duan, Hector Basevi, Hyung Jin Chang, and Ales Leonardis. Pointposenet: Point pose network for robust 6d object pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2824–2833, 2020.

[7] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, and Ales Leonardis. G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4233–4242, 2020.

[8] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021.

[9] Alvaro Collet, Dmitry Berenson, Siddhartha S Srinivasa, and Dave Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *2009 IEEE International Conference on Robotics and Automation*, pages 48–55. IEEE, 2009.

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[11] Yanping Fu, Qingan Yan, Long Yang, Jie Liao, and Chunxia Xiao. Texture mapping for 3d reconstruction with rgb-d sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4645–4653, 2018.

[12] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

[13] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3003–3013, 2021.

[17] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020.

[18] Stefan Hinterstoisser, Cedric Cagniart, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit. Gradient response maps for real-time detection of textureless objects. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):876–888, 2011.

[19] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017.

[20] Taeyeop Lee, Byeong-Uk Lee, Myungchul Kim, and In So Kweon. Category-level metric scale object shape and pose estimation. *IEEE Robotics and Automation Letters*, 6(4):8575–8582, 2021.

[21] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166, 2009.

[22] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.

[23] Haitao Lin, Zichang Liu, Chilam Cheang, Lingwei Zhang, Yanwei Fu, and Xiangyang Xue. Donet: Learning category-level 6d object pose and size estimation from depth observation. *arXiv preprint arXiv:2106.14193*, 2021.

[24] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021.

[25] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015.

[26] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018.

[27] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[28] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[29] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017.

[30] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Feature mapping for learning fast and accurate 3d pose inference from synthetic images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4663–4672, 2018.

[31] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. Instance-and category-level 6d object pose estimation. In *RGB-D Image Analysis and Processing*, pages 243–265. Springer, 2019.

[32] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. A review on object pose recovery: from 3d bounding box detectors to full 6d pose estimators. *Image and Vision Computing*, 96:103898, 2020.

[33] Caner Sahin and Tae-Kyun Kim. Category-level 6d object pose recovery in depth images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[34] Yongzhi Su, Jason Rambach, Nareg Minaskan, Paul Lesur, Alain Pagani, and Didier Stricker. Deep multi-state object pose estimation for augmented reality assembly. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 222–227. IEEE, 2019.

[35] Zhiqiang Sui, Zheming Zhou, Zhen Zeng, and Odest Chadwicke Jenkins. Sum: Sequential scene understanding and manipulation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3281–3288. IEEE, 2017.

[36] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pages 530–546. Springer, 2020.

[37] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991.

[38] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019.

[39] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[40] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021.

[41] Thomas Whelan, Michael Kaess, Hordur Johannsson, Maurice Fallon, John J Leonard, and John McDonald. Real-time large-scale dense rgb-d slam with volumetric fusion. *The International Journal of Robotics Research*, 34(4-5):598–626, 2015.

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.